Assignment 2

CSL7620: Machine Learning

AY 2023-24, Semester – I

Due on: 10/10/2023

M.M: 180 + 10 (Bonus)

General Instructions:

1. Clearly, mention the assumptions you have made, if any.

2. Clearly, report any resources you have used while attempting the assignment.

3. Any submission received in another format or after the deadline will not be evaluated.

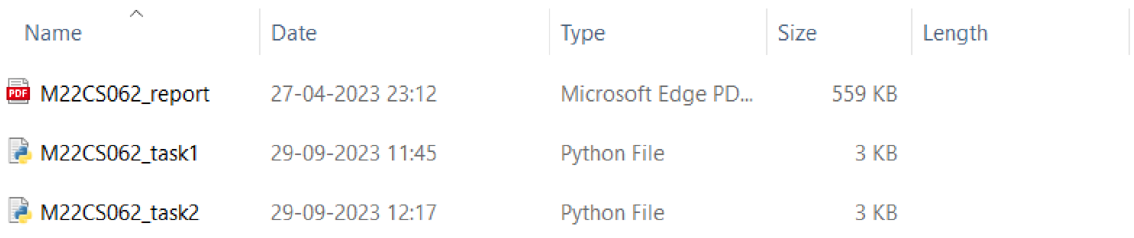
4. Make sure to add references to the resources that you have used while attempting the assignment.

5. Plagiarism of any kind will not be tolerated and will result in zero marks.

6. Select your dataset correctly. If found otherwise, your assignment will not be evaluated.

Submission Guidelines:

1. There should be two .py files named: <roll_no>_task1.py and <roll_no>_task2.py
2. No need to make different py files for subtasks. <roll_no>_task1.py should contain all the subtasks and same for task 2.
3. The .py files must not be named like **<roll no>_task1(1).py**
4. There should be link of **ipnyb** file for each task in the pdf report.
5. **No need to make zip file. Just upload {** <roll_no>_task1.py ,<roll_no>_task2.py and <roll no>_report.pdf**} directly**. So, total 3 files should be there.
6. **Every important result should be reported in the report file.**
7. **In report proper numbering of the tasks/subtasks should be there. Without proper numbering the answer will not be evaluated.**
8. **If the above instructions are not followed there will be penalty.**



Name	Date	Type	Size	Length
M22CS062_report	27-04-2023 23:12	Microsoft Edge PD...	559 KB	
M22CS062_task1	29-09-2023 11:45	Python File	3 KB	
M22CS062_task2	29-09-2023 12:17	Python File	3 KB	

Fig: Example of file names

Dataset link: https://www.kaggle.com/datasets/oddrationale/mnist-in-csv

Task 1:

Imagine if a computer could automatically group similar things together in a massive dataset, without being told what those groups are. That's exactly what K-means clustering does – it's like magic for data!

Have you ever wondered how Netflix suggests movies or how your smartphone sorts your photos? K-means clustering is one of the secrets behind these recommendations and organization.

Picture a scenario where a computer learns to classify animals based on their features without being given any labels – K-means can do that! It's like teaching a computer to think like a biologist.

Let’s try that ourselves: **[75 marks + 5 marks]**

a.) Perform **K-means clustering** on MNIST data **from scratch.** Instead of using Euclidian distance as distance metric use **Cosine Similarity** as distance metric. Clustering should be done in 10, 7, and 4 clusters. **[65 marks]**

b.) Visualize the images getting clustered in different clusters. **[5 marks]**

c.) Please comment on the cluster characteristics. **[5 marks]**

d.) Try to write a python function which finds optimal number of clusters for this dataset? **[Bonus – 5 marks]**

Task 2:

Picture a technique that can simplify complex data while preserving its critical elements, like a magic lens that brings clarity to chaos. PCA (Principal Component Analysis) is the key to unlocking these possibilities, offering a fascinating journey into the world of data transformation and exploration.

Do the followings: **[95 marks + 5 marks]**

a.) Perform PCA on MNIST and then perform GMM clustering. (**Library can be used for SVD and GMM) but PCA should be from scratch**. PCA should be done for 32, 64 and 128 components. Clustering should be done in 10, 7, and 4 clusters. **[80 marks]**

b.) Visualize the images getting clustered in different clusters. **[5 marks]**

c.) Please comment on cluster characteristics and comment with respect to previous task about what difference you see. **[10 marks]**

d.) Can you find the optimal number of components the PCA should choose which covers almost all the necessary patterns in the data? Can you comment on where PCA can fail? [**Bonus 5 marks]**

Report [10 marks]